Jul 12th, 8:30 AM - 8:50 AM

R package MVGHD: causal inference procedure for geographic high-dimensional

Stéphane Bourrelly
University of Lyon III, stephane.bourrelly@univ-lyon3.fr

Pascal Auquier
Aix-Marseille University, pascal.auquier@univmed.fr

International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

R package MVGHD: causal inference procedure for geographic high-dimensional

Stéphane Bourrelly[a], **Pascal Auquier** [b]
[a] *University of Lyon III, UMR 5600 (EVS) CNRS*
(stephane.bourrelly@univ-lyon3.fr)
[b] *Aix-Marseille University EA 3279 (INSERM)*
(pascal.auquier@univmed.fr)

Abstract: Causal statistic inference is a key issue in machine learning. The goal is designing procedures that are able to select relevant subsets of explanatory variables, which might help scientists to better understand the underlying mechanisms behind the studied phenomena. We present a causal statistic inference procedure, especially designed for Geographic High-dimensional Datasets (GHD). The promise of discovering unknown informative factors are as great as the intrinsic learning challenges in these complex datasets, which are more and more common in the fields of environment, health, ecology, epidemiology, geography, agriculture, etc. Firstly, we point out the difference between the variable selection strategies designed for the purpose of "understanding" or "predicting". Then we review the characteristics of the scarce variable selection strategies suitable for the causal statistic inference. Next we highlight the complexity of GHD; through the one included in the presented R package. This latter was created with the objective of better understanding the health impacts of 63 environmental factors, from the hundreds of sources of the so-called "French environmental big data" and the medical database: LEA. Indeed, at geographical scales the variety of available data allows the study of unexplored environmental factors (e.g. chronic exposure to trace metals or radiation). However, the GIS aggregations performed to create the final spatial indicators is very time consuming and decrease the accuracy of sources. Therefore the studied phenomenon (e.g. the morbidity) is usually represented both by a numerical and a multiclass spatial indicator. In addition, the potential explanatory spatial indicators (e.g. environmental factors) are also qualitative or quantitative and more or less correlated. Moreover, they are known in a very low number of spatial units. Secondly, from the previous considerations we explain why at present only the heuristic variable selection strategies based on Random Forests (FR) are able to handle the GHD. Then we present step by step the backgrounds of the causal inference procedure: MVGHD, through the convenient (beta) functions of the R package applied to this eco-epidemiological GHD. For each step we explain how to run the procedure in order to select relevant subsets of explanatory variables. 'mvg.tune(.)' optimizes the parameters of RF through a trade-off between statistical accuracy and computational time. 'mvg.select(.)' selects and compares the subset of explanatory spatial indicators, by the two different and customised variable selection strategies. 'mvg.estimate(.)' assesses the significance of results, by performing cross-validation techniques. 'mvg.display(.)' provides statistical summaries, charts and maps, so to help interpreting the results. Finally we conclude on the strengths and limits of the understanding thus gained on the role played by the combined effects of environmental factors on health risks.

***Keywords**:* Variable selection, random forests, high-dimensional dataset, lattice.